Parsimonious mixed models

Douglas Bates[a], Reinhold Kliegl[b], Shravan Vasishth[b], and Harald Baayen[c]
[a] University of Wisconsin-Madison, USA
[b] University of Potsdam, Germany
[c] University of Tübingen, Germany

Running Head:

Corresponding author:
R. Harald Baayen
Seminar für Sprachwissenschaft
Eberhard Karls University Tübingen
Wilhelmstrasse 19
Tübingen
e-mail: harald.baayen@uni-tuebingen.de

Abstract

The analysis of experimental data with mixed-effects models requires decisions about the specification of the appropriate random-effects structure. Recently, Barr et al. (2013) recommended fitting 'maximal' models with all possible random effect components included. Estimation of maximal models, however, may not converge. We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling with uninformative or weakly informative priors. Importantly, even under convergence, overparameterization may lead to uninterpretable models. We provide diagnostic tools for detecting overparameterization and guiding model simplification. Finally, we clarify that the simulations on which Barr et al. base their recommendations are atypical for real data. A detailed example is provided of how subject-related attentional fluctuation across trials may further qualify statistical inferences about fixed effects, and of how such nonlinear effects can be accommodated within the mixed-effects modeling framework.

Keywords: linear mixed models, generalized additive mixed models, model selection, crossed random effects, model simplicity

1 Introduction

During the last ten years, there has been a significant change in how psycholinguistic experiments are analyzed when both subjects and items are included as random factor, specifically a change from analyses of variance to linear mixed models (LMM), with Baayen et al. (2008) providing a first major introduction. Within psychology this change is spreading to other areas, most notably to personality and social psychology (Judd et al., 2012; Westfall et al., 2014). There are a number of reasons for this change. One particularly attractive feature has been that, with LMMs, statistical inference about experimental effects and interactions no longer needs separate analyses of variance, one for subjects and one for items (Clark, 1973; Forster and Dickinson, 1976), but can be carried out within a single coherent framework.

This benefit in coherence comes at some cost. An important part of analyzing experimental data with mixed-effects models is the selection of the proper random-effects structure. In principle, LMMs not only consider variance between subjects and between items in the mean of the dependent variable (i.e., random intercepts), but also variance between subjects and between items for all main effects and interactions (i.e., random slopes) as well as correlations between intercepts and slopes. Let us illustrate the basic point with the most simple model with a two-level within-subject experimental manipulation and subject as the only random factor, using the formula notation of the `lme4` package for R described in Bates et al. (2015):

```
Y ~ 1 + A + (1|Subject)
```

This model may support the fixed effect of the within-subject factor `A`. However, if the effect of `A` (i.e., the difference between the two experimental conditions) differs reliably between subjects, uncertainty about `A` may be so substantial that the main effect of `A` is no longer significant in a model allowing for random slopes for `A`:

```
Y ~ 1 + A + (1+A|Subject)
```

For the assessment of the significance of the experimental manipulation, it is therefore essential to examine its fixed effect in the presence of the corresponding random slopes (see, e.g., Pinheiro and Bates, 2000; Baayen, 2008).

For models with several fixed factors (e.g., experimental manipulations) and several random factors (e.g., subjects and items), the question of how to choose the appropriate random-effects structure becomes substantially more complex. Schielzeth and Forstmeier (2009) demonstrated that both random intercepts and random slopes need to be considered in LMMs to guard against anti-conservative conclusions (i.e., accepting an experimental effect more frequently as significant than warranted by the data). Inclusion of random slopes also reduces residual variance and increases statistical power for the detection of between-subject and between-item effects. Addressing this same issue on the basis of simulation studies, Barr et al. (2013) recommended, with some qualifications and caveats, that in order to avoid anti-conservative conclusions, mixed-effects models should be specified with a maximal random-effects structure. For instance, for a 2×2 within-subject and within-item factorial design, their recommendation is to fit the 'maximal' model

```
Y ~ 1 + A + B + A:B + (1 + A + B + A:B|subject) + (1 + A + B + A:B|item)
```

Unfortunately, often the amount of information in the data (e.g., the number subjects or the number of observations per subject, contingent also on the within-group standard deviation) is not sufficient for the reliable estimation of all the parameters available in principle. One reason is that the number of model parameters grows quadratically with the number of variance components (i.e., for n variance components related to a random factor we obtain maximally $n(n+1)/2$ model parameters, ignoring fixed effects). So the question is: How can we arrive at a "defensible" model given the available amount of information in the data?

Barr et al. (2013) argue that failure to specify a maximal random effects structure amounts to violating the compound symmetry assumption in classical analysis of variance (ANOVA). Their demand to keep models (ideally) maximal is of relevance in the experimental psychological or psycholinguistic context, because these fields of research have a very strong affiliation with frequentist statistics. Correcting alpha levels for multiple tests (e.g., FDR, Bonferroni) and adjusting degrees of freedom for violations of assumptions about sphericity/compound symmetry (e.g., Greenhouse-Geisser, Huyn-Feldt) are part of the routine practice for establishing the significance of effects.

The recommendations by Barr et al. (2013) to keep random-effects structure maximal has increased awareness of the importance of random effects structure other than simple random intercepts. However, as models become more complex and given the typical number of subjects and items in psycholinguistic experiment, it becomes increasingly more difficult to estimate the model's parameters. Indeed, often the estimation algorithm may not even converge anymore.

Barr et al. (2013) base their recommendations on simulation studies comparing different procedures for model selection. One could start with a minimal model, and incrementally add predictors and random-effects structure. Or one could start with a maximal model, and incrementally simplify it. Combinations of forward and backward selection are also possible. Interestingly, Barr et al. observed that maximal models, i.e., from the set of estimable models those models that have maximal random effects structure, yield the best results.

The present study has three goals. First, we explain that failure of convergence of model estimation is typically not a consequence of a suboptimal estimation algorithm, but rather an indicator of a model specification that is too complex to be properly supported by the data. In the absence of convergence, parameter estimates of overparameterized LMMs are not interpretable.

Second, we introduce diagnostic procedures for simplifying models with overspecified random-effects structure. We do this within the LMM framework relying (1) on a principal component

analysis (PCA) of the random effects structure to determine the number of variance components and correlation parameters supported by the data and (2), once the most complex model supported by the data is identified, on comparisons of goodness of fit of nested models with likelihood ratio tests (LRTs).

We show for two experiments here and several other experiments discussed in detail in the supplementary materials — the 'vignettes' of the `RePsychLing` package[1] — that the same conclusions are reached irrespective of whether parameter estimation is carried out using maximum likelihood estimation (implemented in the `lme4` package; (Bates et al., 2014a)) or with a Bayesian hierarchical linear model using uninformative or mildly informative priors (Gelman et al., 2014). We stress the importance of bringing the information in the data and factorial LMM model complexity in agreement, which typically leads to more parsimonious models with a smaller than theoretically possible number of model parameters.

Third, the simulations on which Barr et al. (2013) base their recommendation are both too simple and too atypical for real data to support the claim that maximal random effects structure is desirable. We discuss an example illustrating that bringing into the analysis hidden, typically nonlinear, processes, often related to the human factor, affords more precise inference about both random and fixed effects.

2 Parameter estimation in mixed models

Following the publication of Barr et al. (2013) containing the advice to "keep it maximal" when formulating an LMM for confirmatory analysis, the frequency of reports and queries related to failure of a model fit to converge has increased, for example, in the discussion list for the `lme4` (Bates et al., 2014b) package for R. Although LMMs and generalized linear mixed-effects models (GLMMs) are versatile tools for modeling the variability in observed responses and attributing parts of this variability to different sources, like any statistical modeling technique they have their limitations. Knowledge of these limitations is important in ensuring appropriate usage.

As the complexity of the model increases, so does the difficulty of the optimization problem. For LMMs, aside from intercept and fixed effects, the parameters being estimated represent variances and covariances, which are typically much more difficult to estimate than regression coefficients. The parameters in GLMMs are even more complicated.

When there is more than one experimental factor, say in a factorial design, the number of parameters to be estimated explodes. For example, a maximal model with three experimental within-subject and within-items factors with two levels each in a full factorial design incorporating the seven main effects and interactions plus the intercept with random effects for subject and item, would require estimation of eight fixed-effects coefficients and 73 variance-covariance parameters. The online supplement to Barr et al. (2013) fits such a model to data from an experiment described in Kronmüller and Barr (2007). We also present a reanalysis of this experiment below.

To anticipate the main result, it is simply not realistic to try to fit this number of highly nonlinear parameters given the number of subjects and items in this experiment. Almost unfortunately, the software does indeed converge to parameter estimates but these estimates correspond to degenerate or singular covariance matrices, in which some linear combinations of the random effects are estimated to having no variability. This corresponds to estimates of zero random-effects variance in a model with random-intercepts only or a correlation of ± 1 in a model with correlated random intercepts and slopes. However, already a three-by-three correlation matrix will not usually show boundary values like these, even when it is singular. In summary, the parameters representing variances and

[1] Available from https://github.com/dmbates/RePsychLing

covariances are constrained in complicated ways. In overparameterized models, convergence can occur on the boundary, corresponding to models with singular variance-covariance matrices for random effects. This can have serious, adverse consequences for inference; for example, due to an overparameterization of the maximal LMM, Kliegl et al. (2011) wrongly interpreted an LMM correlation parameter as providing much more evidence than the corresponding within-subject correlation for the correlation of two experimental effects.

In a linear mixed model incorporating vector-valued random effects, say by-subject random effects for intercept and for slope, the variance component parameters determine a variance-covariance matrix for these random effects. As described in Bates et al. (2015), the parameters used in fitting the model are the entries in the Cholesky factor (Λ) of the relative variance-covariance matrix of the unconditional distribution of the random effects. The parameter vector (θ) for this model are the values on and below the diagonal of a lower triangular Cholesky factor. The θ vector elements fill the lower triangular matrix in column major order. The relative covariance matrix for the random effects is $\Lambda\Lambda'$. To reproduce the covariance matrix, $\Lambda\Lambda'$ must be scaled by s^2.

When one or more columns of the Cholesky factor Λ are zero vectors, Λ is rank-deficient: The linear subspace formed by all possible linear combinations of the columns is of reduced dimensionality compared to the dimensionality of Λ. The random-effects vectors that can be generated from this fitted model must lie in this lower-dimensional subspace. That is, there will be no variability in one or more directions of the space of random effects.

The `RePsychLing` package provides a new function, `rePCA` (which may become part of a future release of `lme4`) that enables the analyst to probe models fitted with `lmer` for rank deficiency. The `rePCA` (random-effects Principal Components Analysis) function takes an object of class `lmerMod` (i.e. a model fit by `lmer`) and produces a list of principal component (`prcomp`) objects, one for each grouping factor. These principal component objects can be summarized and visualized (by means of scree plots), exactly as any other principal component object generated by the `prcomp` function of `R`.

Principal components analysis of the estimated covariance matrices for the random effects in a linear mixed model allows for simple assessment of the dimensionality of the random effects distribution. As illustrated below and in the vignettes in the `RePsychLing` package, the maximal model in many analyses of data from Psychology and Linguistics experiments, is almost always shown by this analysis to be degenerate.

3 Iterative reduction of model complexity

3.1 Overview

In this section, we assume that the researcher has hypotheses about main effects and (some) interactions, but that he/she has no specific expectations about variance components or correlation parameters. In other words, we assume that the experimental hypotheses relate to the fixed effects, not the random-effects structure. In hypothesis testing, usually the primary reason for dealing with the random-effects structure is to obtain as powerful tests as justified of the fixed effects. Therefore, it is reasonable to remove variance components/correlation parameters from the model if they are not supported by the data. If there are specific expectations about, say, a correlation parameter, it makes sense to include it in the model (as well as the related variance components). We also assume the standard situation whereby potential numeric within-subject or within-item covariates are not under consideration, but we will return to this case in the section entitled *Hidden complexities*.

In a factorial experiment, the maximum number of variance-covariance parameters to be estimated

for each random factor is

$$\frac{(\text{product of within-factor levels}) \times (\text{product of within-factor levels} + 1)}{2}.$$

For example, a 2×2 within-factor design will have $(2 \times 2) \times ((2 \times 2) + 1)/2 = 10$ parameters in the variance-covariance matrix for each random effect (commonly, subject and item), and a $2 \times 2 \times 2$ within-factor design will have 36 parameters. Additional model parameters are required for the fixed effects intercept, main effects, interactions, and for the residual variance. It is reasonable to start by attempting to fit a maximal LMM. If this model converges within reasonable time, several steps can be taken to check the possibility of an iterative reduction of model complexity in order to arrive at a parsimonious LMM.

First, it is worth checking whether we can reduce the dimensionality of the variance-covariance matrices assumed in a maximal LMM. The number of principal components that cumulatively account for 100% of the variance is a reasonably stringent criterion for settling on the reduced dimensionality. This can be achieved by performing PCA using the `rePCA()` function on the fitted maximal LMM (see the section above entitled *Parameter estimation in mixed models*).

Second, after we have determined the number of dimensions supported by the data, we can eliminate variance components from the LMM, following the standard statistical principle with respect to interactions and main effects: variance components of higher-order interactions should generally be taken out of the model before lower-order terms nested under them. Frequently, in the end, this leads also to the elimination of variance components of main effects. The reduced model may be submitted again to a PCA to check the dimensionality of the random-effects structure.

Third, we can check whether forcing to zero the correlation parameters of the reduced LMM significantly decreases the goodness of fit according to a likelihood ratio test (LRT), possibly also taking into account changes in AIC and BIC. Obviously, if the goodness of fit does not change from the reduced model to the zero-correlation-parameter (ZCP) model, we do not have reliable evidence that the correlation parameters are different from zero. Importantly, this does not mean that the correlations are zero, only that we do not have enough evidence for them being different from zero for the current data; absence of evidence is not evidence of absence. Also, note that the estimated value of the correlation parameters depends on the choice of contrasts for the experimental factors. For example, treatment contrasts and sum contrasts may lead to models with a very different random-effect structure (see http://www.rpubs.com/Reinhold/22193; this is also available as a vignette in the RePsychLing package). It is also conceivable that correlation parameters are zero for only one of several random factors. The new '||' syntax of `lmer()` is very convenient for specifying such ZCP LMMs. However, as illustrated in the same vignette, there are a few constraints relating to general R-formula syntax one needs to know about. In general, the '||' syntax works (currently) as expected only for LMMs after converting factor-based to vector-valued random-effects structures.

Fourth, we may also want to check whether all the variance components of an identified model are necessary. Taking out one term at a time and checking again whether there is a significant drop in goodness of fit, allows us to identify variance components that are not supported by the data. Again, removing such terms does not mean that the variance is zero, only that we have no evidence of it being significantly different from zero.

Fifth, after removing non-significant variance components, we may want to recheck whether the goodness of fit of the iteratively reduced model increases if it is extended with correlation parameters. A reliable variance parameter (i.e., a variance parameter contributing significantly to the goodness of fit according to a likelihood-ratio test) is a necessary condition for estimating correlation parameters associated with this variance component. In other words, we expect to find statistically significant correlation parameters only if the related variance components are statistically

significant by themselves.

Finally, as a special case, the iterative reduction of model complexity described above assumed that there was a solution for the maximal model. With complex experiments, however, it may happen that the maximal model does not converge to a solution (e.g., there are warnings that no solution was found) or that the solution is obviously degenerate (i.e., variance components or correlation parameters are estimated at their boundaries of 0 or $+1/-1$, respectively). In this case, a first step could be to check the dimensionality of the zero-correlation parameter model with a PCA. Obviously, with complex experiments the switch from maximal to zero-correlation parameter model will yield the largest simplification. As already mentioned, whenever one switches from maximal to zero-correlation parameter models, it is very important to have a clear understanding of the contrast specifications chosen for the experimental factors.

In the following two sections, we describe iterative reductions of LMM complexity for two experiments using the above checks. The data for the first example are from an experiment on pragmatic comprehension of instructions (Kronmüller and Barr (2007), Exp. 2; reanalyzed with an LMM in Barr et al. (2013)). The second data set is from a visual-attention experiment (Kliegl et al., 2015), following up a different report with the overparameterized model (Kliegl et al., 2011). Detailed reports (including R code) for these analyses are described in vignettes, along with four additional examples. We emphasize that we do not claim that our illustrations are the only way to carry out these analyses, but the strategy outlined above has yielded satisfactory results with all data sets we have analyzed so far. There is no cook-book substitute for theoretical considerations and developing statistical understanding. Each data-set deserves the exercise of judgement on part of the researcher.

3.2 Reanalysis of Kronmüller and Barr (2007)

Here we apply the iterative reduction of LMM complexity to truncated response times of a $2 \times 2 \times 2$ factorial psycholinguistic experiment (Kronmüller and Barr, 2007). This is their Exp. 2, reanalyzed with an LMM in Barr et al. (2013). The data are from 56 subjects who responded to 32 items. Specifically, subjects had to select one of several objects presented on a monitor with a cursor. The manipulations involved (1) auditory instructions that maintained or broke a precedent of reference for the objects established over prior trials, (2) with the instruction being presented by the speaker who established the precedent (i.e., an old speaker) or a new speaker, and (3) whether the task had to be performed without or with a cognitive load consisting of six random digits. All factors were varied within subjects and within items. There were main effects of Load (L), Speaker (S), and Precedent (P); none of the interactions were significant. Although standard errors of fixed-effect coefficents varied slightly across models, our reanalyses afforded the same statistical inference about the experimental manipulations as the original article, irrespective of LMM specification (see Figure 1 a comparison of fixed effects of maximal and parsimonious LMMs). The purpose of the analysis is to illustrate an assessment of model complexity as far as variance components and correlation parameters are concerned, neither of which were the focus of the original publication.

Maximal linear mixed model. A full factorial model in the fixed-effects can be described by the formula , 1 + S + P + C + SP + SC + PC + SPC. Barr et al. (2013) analyzed Kronmüller and Barr (2007, Exp. 2) with the maximal model for this design comprising 16 variance components (eight each for the random factors `SubjID` and `ItemID`, respectively). The model took 39,004 iterations to converge, but produces what look like reasonable parameter estimates (i.e., no variance components with estimates close to zero; no correlation parameters with values close to ± 1). The slow convergence is due to the total of $2 \times 36 = 72$ parameters in the optimization of the random-effects part (ignoring

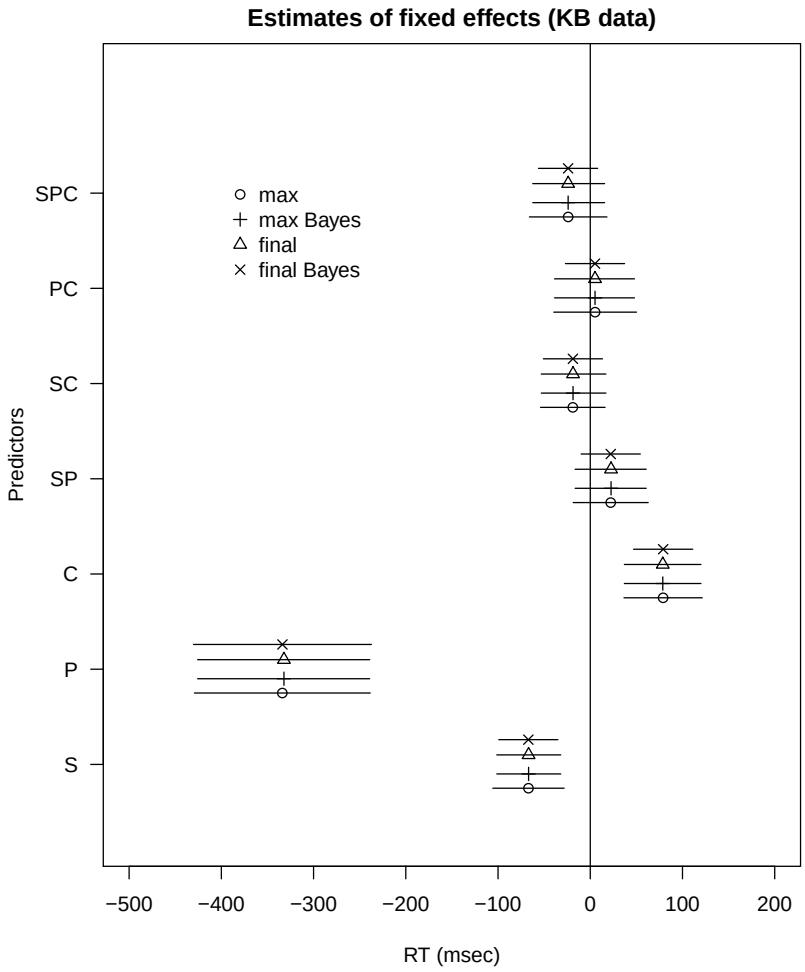

Figure 1: The estimates and 95% confidence intervals for the fixed effects in the maximal and final models of the Kronmüller and Barr 2007 data. Also shown are estimates and 95% credible intervals from a maximal Bayesian hierarchical linear model.

the eight fixed-effect parameters and the residual variance, which do not contribute much to the computational load). Figures 1 and 2 display the fixed effects and variance components of the maximal model, along with other estimates, discussed below. The correlations of subject and item random effects are shown in Figure 3.

Considering that there are only 56 subjects and 32 items, it is quite optimistic to expect to estimate 36 covariance parameters for `SubjID` and another 36 for `ItemID`. A principal components analysis of the variance-covariance matrices for subject and item random effects returns eight principal components, along with the cumulative proportions of variance explained (see Table 1). For subject random effects, four dimensions are sufficient to account for 100% of the variance explained; and for items, five dimensions suffice.

Thus, the maximal model is clearly too complex. In the following paragraphs, we illustrate our iterative method that reduces model complexity to arrive at an optimal LMM for this experiment. We will not report the intermediate results here, but they are available in the vignettes, along with the R code. We qualify this procedure at the outset: We do not claim that this is the only way to proceed, but the strategy has consistently yielded satisfactory results for all data sets we have

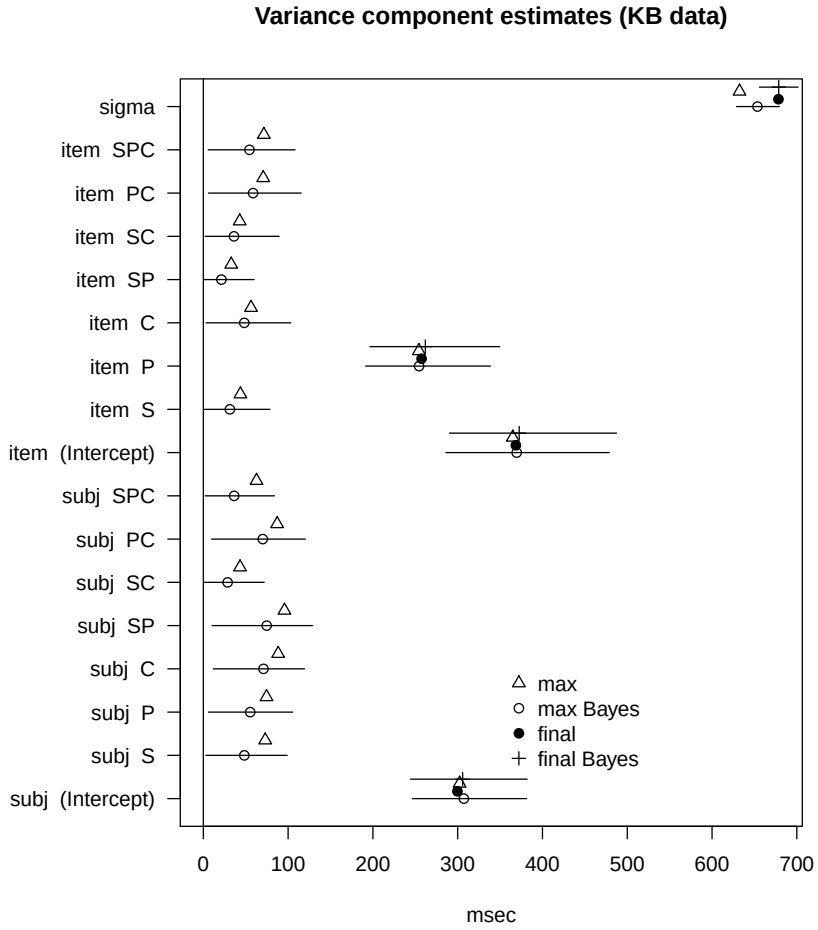

Figure 2: The estimates of the variance components (standard deviations) from the maximal and final models of the KB data. Also shown are the estimates and 95% credible intervals of the maximal Bayesian hierarchical model.

		1	2	3	4	5	6	7	8
subject	cum. prop.	0.73	0.85	0.94	1.00	1.00	1.00	1.00	1.00
item	cum. prop.	0.79	0.94	0.97	0.99	1.00	1.00	1.00	1.00

Table 1: The cumulative proportion of variance explained for the subject and item random effects in the maximal model for the Kronmüller and Barr 2007 data. Principal components analysis was used to compute the cumulative proportion of variance explained.

examined so far.

Zero-correlation-parameter linear mixed model. As a first step toward model reduction, we start with a model including all 16 variance components, but no correlation parameters. The PCA of this model shows that 12 out of 16 dimensions suffice for capturing 100% of the variance explained. This suggests that the model is still too complex.

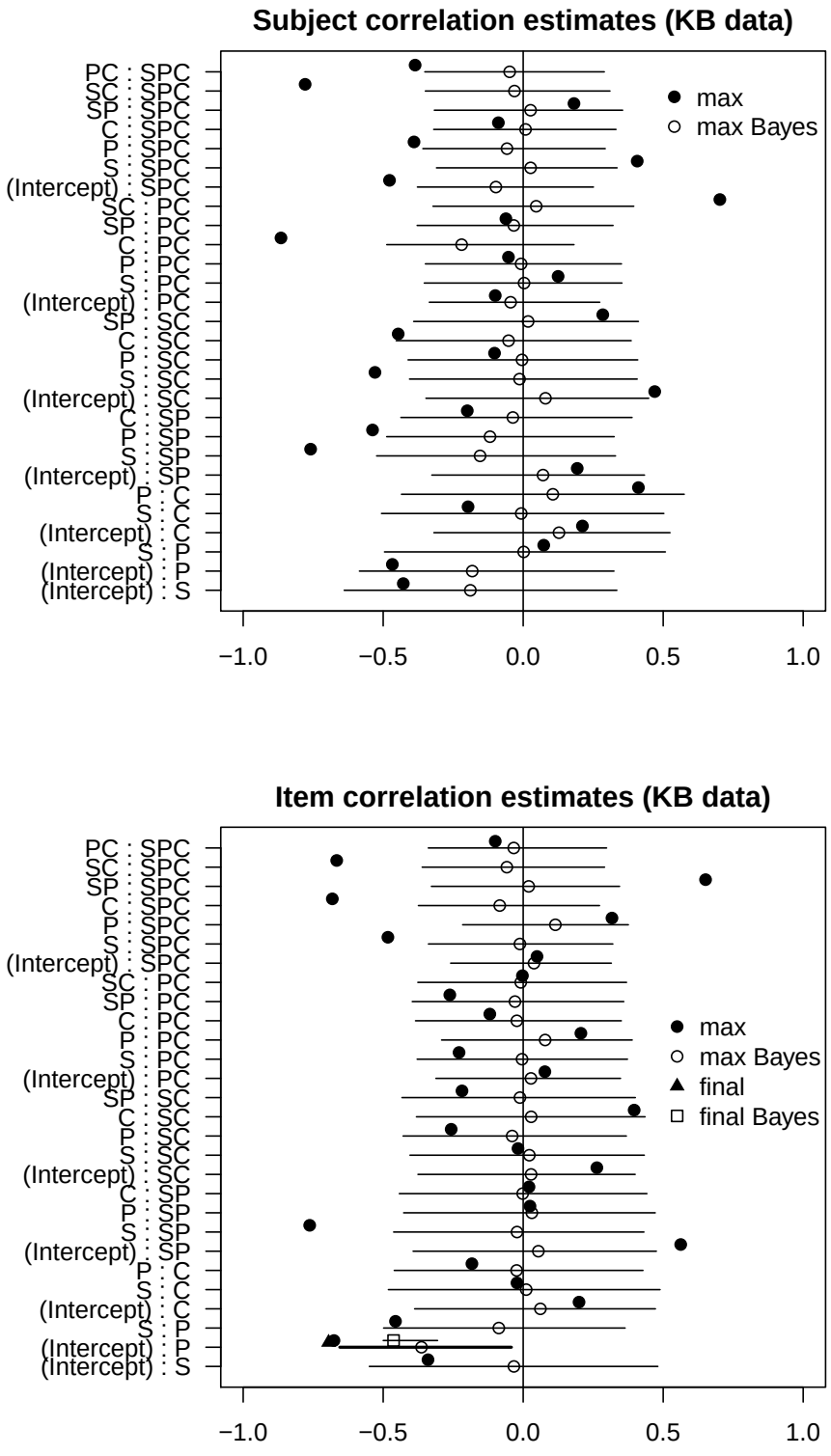

Figure 3: The estimates of the subject and item random effect correlations in the `lme4` and Bayesian maximal models of the Kronmüller and Barr 2007 data. The Bayesian estimates also have 95% credible intervals.

Dropping variance components to achieve model identification. A second step toward model reduction could be to remove variance components to achieve model identification. Starting with the smallest variance component (or a set of them) this step can be repeated until the PCA no longer suggests overidentification. For the present case, variance components 5 and 7 for `SubjID` and 1 and 4 for `ItemID` are estimated with zero or close to zero values. We refit the LMM without these variance components. The PCA for this LMM estimates 12 non-zero variance components.

Dropping non-significant variance components. In the third step we attempt to simplify the random-effects structure of the identified LMM with likelihood ratio tests. For example, the two smallest variance components account for less than 1% of the variance. We iteratively remove variance components, starting with dropping the highest-order interaction term SPC. Moving on to tests of two-factor interactions, we end up with an LMM comprising only varying intercepts for subject and items and the item-related variance component for P. Looking back to the maximally identified LMM, we see that these are exactly the three variance components with clearly larger standard-deviation estimates (> 249) compared to the other standard-deviation estimates (> 64). There is no significant loss of goodness of fit when we remove nine variance components identified this way; $\chi^2_9 = 11.1$, $p = .27$. However, removal of any of three remaining variance components significantly reduces the goodness of fit.

Extending the reduced LMM with a correlation parameter. Inclusion of the correlation parameter between item-related intercept and the precedence effect (P) for this model significantly improves the goodness of fit with the correlation parameter estimated at -0.69; $\chi^2_1 = 16.3$, $p < .01$. Thus, there is evidence for reliable differences between items in the precedence effect. The variance components and correlation parameter for this final LMM are displayed in Figure 2.[2]

Summary. In our opinion, the final model we settled on is the *optimal* LMM for the data of this experiment. To summarize our general strategy: (1) we started with a maximal model; (2) then, we fit a zero-correlation model; (3) next, we removed variance components until the likelihood ratio test showed no further improvement; and (4) finally, we added correlation parameters for the remaining variance components. Principal components analysis was used throughout to check the dimensionality for the respective intermediate models. This approach worked quite well in the present case. Indeed, we also reanalyzed three additional experiments reported in the supplement of Barr et al. (2013). As documented in the `RePsychLing` package accompanying the present article, in each case, the maximal LMM was too complex for the information provided by the experimental data. In each case, the data supported only a very sparse random-effects structure beyond varying intercepts for subjects and items. Fortunately and interestingly, none of the analyses impacted the statistical inference about fixed effects in these experiments. Obviously, this cannot be ruled out in general. If authors adhere to a strict criterion for significance, such as $p < .05$ suitably adjusted for multiple comparisons, there is always a chance that a t-value will fall above or below the criterion across different versions of an LMM.

[2] Incidentally, although we consider it questionable to compare non-identified and identified models with an LRT, we want to mention that there is no significant difference in goodness of fit between the final LMM and the maximal model we started with. The final number of principal components suggested by the final model is actually smaller than suggested by the initial PCA of maximal model.

3.3 An alternative analysis of Kronmüller and Barr (2007) using a Bayesian LMM

We also show that similar conclusions can be reached if we fit a Bayesian linear mixed model (Gelman et al., 2014) instead of the frequentist model discussed above using `lme4`. Presenting the Bayesian estimates corresponding to the maximal linear mixed model presented above provides an independent validation of the conclusion that a simpler model has better motivation.

We fit a linear mixed model to the Kronmüller and Barr data using `rstan` (Stan Development Team, 2014a). In a Bayesian linear mixed model, all parameters have a prior distribution defined over them; this is in contrast to the frequentist approach, where each parameter is assumed to have a fixed but unknown value. Defining a prior distribution over each parameter expresses the researcher's belief about possible values that the parameter can take, before any data is considered. For example, in the Bayesian formulation, an `lme4` style model specification such as

```
Y ~ 1 + A + (1|Subject)
```

we could specify that our prior belief about the parameter, call it β_1, expressing an effect of A is that it has a normal distribution with mean 0 and some large variance σ^2. We can write this as

$$\beta_1 \sim \text{Normal}(0, \sigma^2). \tag{1}$$

Such a prior expresses the belief that, in the absence of any data, the mean is assumed to be zero, but the large variance expresses uncertainty about this belief. Using computational methods available in `rstan`, this prior specification can be combined with the data available to derive a posterior distribution for each parameter, including the random effects variance components and the correlations between variance components. The posterior distribution of each parameter is effectively a compromise between the prior and the data, and expresses our revised belief about the parameter's distribution after the data are taken into account. If there is strong evidence from the data that the mean of its distribution is different from zero, the mean of the posterior distribution will reflect this. If there is only weak or no evidence from the data—either due to there being too little data or because the mean from the data is near zero—that the parameter has a mean different from zero, then the prior mean of zero will dominate in determining the parameter's posterior distribution.

The end-product of a Bayesian linear mixed model is always a posterior distribution for each parameter in the model. Thus, we can plot the 95% credible interval for each parameter; this interval tells us the range over which we can be 95% certain that the true value of the parameter lies, given the specific data that we have. Contrast this with the 95% confidence interval (CI), which represents one of hypothethically computed CIs over repeated experiments, where 95% of those hypothetical CIs would contain the true value of the parameter. Note, however, that for relatively large data-sets such as the present one, the credible interval and confidence interval for the fixed effects will generally be identical (see Figure 1).

We fit a linear mixed model to the Kronmüller and Barr data with normal priors on the fixed effects parameters, and a so-called LKJ prior on the correlation matrices of the subject and item random effects (Stan Development Team, 2014b). The LKJ prior assumes that the correlations are zero (with some uncertainty associated with this belief); if there is evidence in the data for a non-zero correlation, the posterior distribution of the correlation parameter will be shifted away from zero. For the standard deviations, we defined a uniform prior with a bound 0. This expresses the prior belief that we have no strong beliefs about the standard deviation, but we know that it cannot be less than 0 (our analysis does not depend on using this prior). For a detailed specification of the model, see the `RePsychLing` package.

The posterior distributions of all parameters for the 'maximal' model, along with their 95% credible intervals, are shown in Figures 1, 2, and 3. The Bayesian analysis shows two important things. First, the estimates of the fixed effects in the `lme4` model and the Bayesian model are nearly identical. This shows that the 'maximal' LMM fit using `lme4` is essentially equivalent to fitting a Bayesian LMM with regularizing priors of the sort described above. Second, the relevant variance component parameters that were identified above using principal components analysis (PCA) and likelihood ratio tests (LRTs) are exactly the parameters that clearly dominate in the Bayesian analysis; see Figures 2 and 3. Specifically, any variance component excluded in the `lme4`-based analysis using PCA and LRTs has, in the Bayesian analysis, a posterior credible interval that includes zero; and any correlation parameter excluded in the `lme4`-based analysis has, in the Bayesian model, a credible interval that spans zero. In other words, when we approach the analysis from the perspective of Bayesian modeling, we also find that there is no evidence in the data that the relevant parameters have values that are different from zero. These parameters should be excluded from the model on grounds of parsimony. For comparison, we also present Bayesian estimates of the final model with a reduced number of variance components (Figures 1-3).

3.4 Reanalysis of Kliegl et al. (2015)

As a second demonstration that linear mixed models with a maximal random-effect structure may be asking too much, we re-analyze data from a visual-attention experiment (Kliegl et al., 2015), following up a published experiment (Kliegl et al., 2011) with an (unfortunately) overidentified LMM, as shown in the vignette for the KWDYZ data in the RePsychLing package. The experiment shows that validly cued targets on a monitor are detected faster than invalidly cued ones (i.e., spatial cueing effect; Posner (1980)) and that targets presented at the opposite end of a rectangle at which the cue had occurred are detected faster than targets presented at a different rectangle but with the same physical distance (object-based effect; Egly et al. (1994)). Different from earlier research, the two rectangles were not only presented in cardinal orientation (i.e., in horizontal or vertical orientation), but also diagonally (45° left or 45° right). This manipulation afforded a follow up of a hypothesis that attention can be shifted faster diagonally than vertically or horizontally across the screen (Kliegl et al., 2011; Zhou et al., 2006)). Finally, data are from two groups of subjects, one group had to detect small targets and the other large targets. The experiment is a follow-up to Kliegl et al. (2011) who used only small targets and only cardinal orientations for rectangles. For the interpretation of the fixed effects, we refer to Kliegl et al. (2015). Again, the different model specifications reported in this section were of no consequence for the significance or interpretation of fixed effects, but they led to inappropriate conclusions about the correlations between variance components. We focus here on exploring the random-effect structure for these data.

Maximal linear mixed model. We start with the maximal linear mixed model including all possible variance components and correlation parameters associated with the four within-subject contrasts in the random-effects structure. Note that there are no interactions between the three contrasts associated with the four levels of the cue-target relation factor. Also, as factor size was manipulated between subjects, this contrast does not appear in the random-effect structure. Thus, the random-effect structure comprises eight variance components (i.e., the intercept estimating the grand mean of log reaction time, the three contrasts for the four types of cue-target relation, the contrast for the orientation factor, and three interactions) and 28 correlation parameters ($8 \times 7/2$)–also a very complex model.

The maximal model converges with a warning:

```
maxfun < 10 *length(par)^2 is not recommended
```

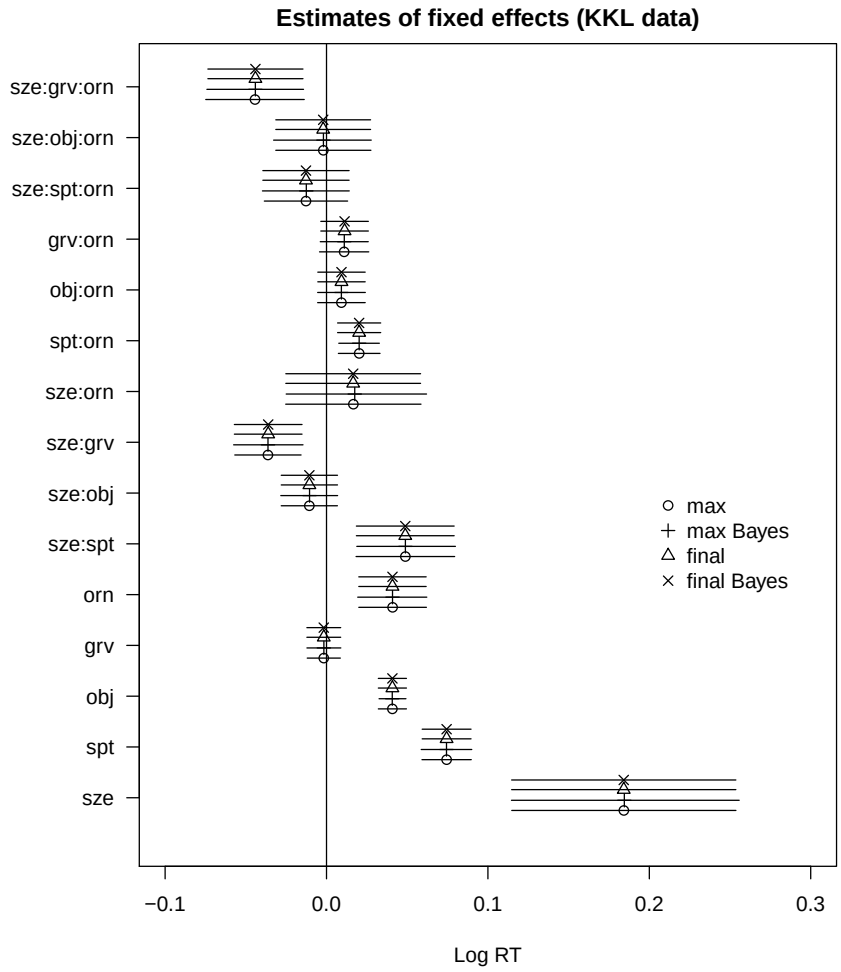

Figure 4: The estimates and 95% confidence intervals for the fixed effects in the maximal and final models of the Kliegl et al 2015 data. Also shown are estimates and 95% credible intervals from maximal and final Bayesian hierarchical linear models.

This suggests that we may be asking too much of these data. Nevertheless, at a first glance, model parameters look reasonable. As shown in Figure 5, none of the eight variance components are estimated at zero and none of the 28 correlation parameters are at the boundary (i.e., none assume values of $+1$ or -1). The PCA, however, indicates that the maximal model is overparameterized: two dimensions contribute 0% variance explained.

Zero-correlation-parameter linear mixed model. The problem of overidentification persists in the zero-correlation parameter model (ZCP LMM) . In the PCA, we still have only seven of eight non-zero components and one of them accounts for less that 1% of the variance. Thus, the ZCP LMM still is too complex for the information contained in the data of this experiment.

Dropping variance components to achieve model identification. The estimates of variance components suggest that there is very little reliable variance associated with the interaction between object and orientation contrasts and for the interaction between gravitation and orientation. Dropping

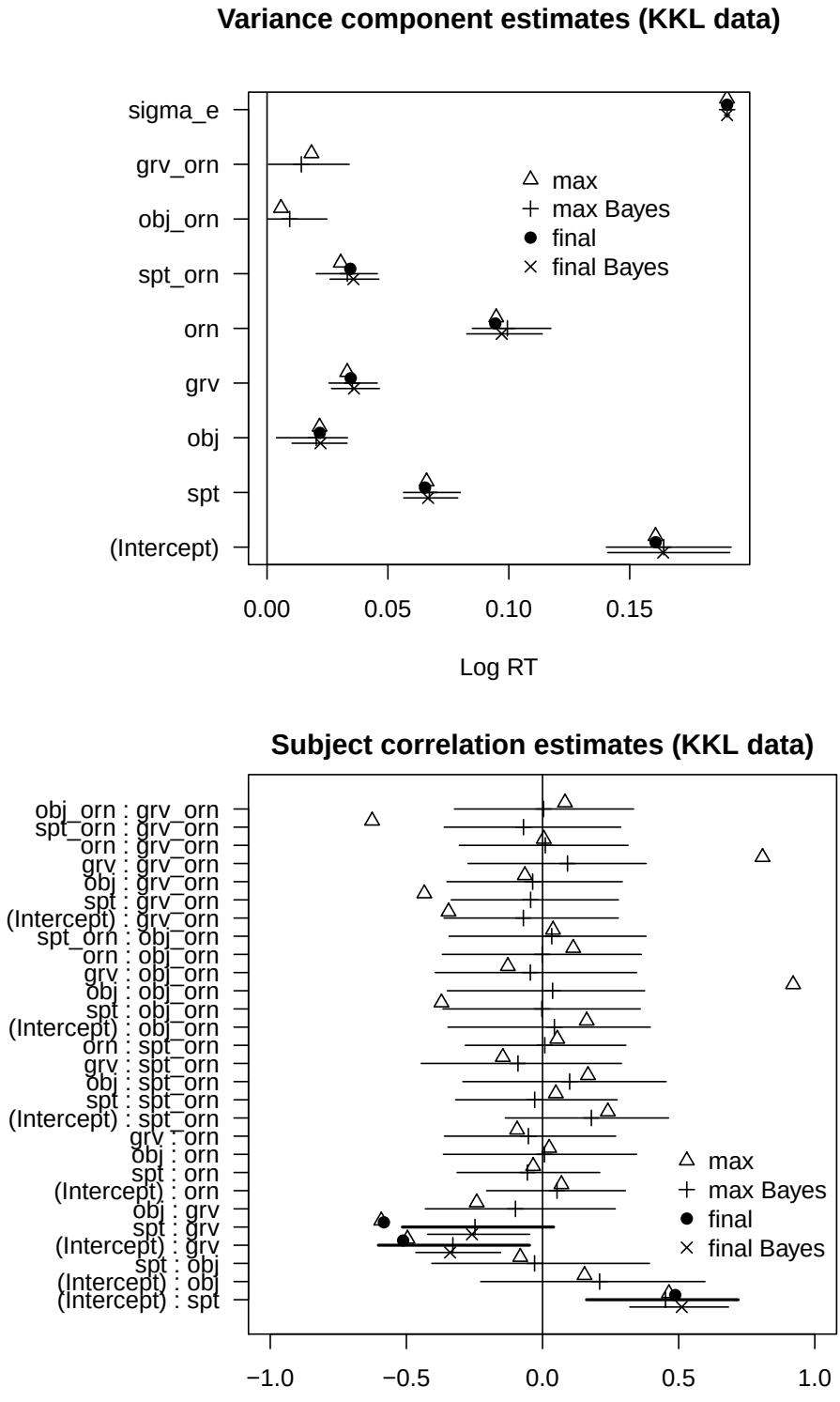

Figure 5: Top: The standard deviation estimates from the maximal and final models of the KKL data, and estimates and 95% credible intervals of the maximal and final Bayesian hierarchical models. Bottom: Correlations of subject random effects in the maximal and final models of the KKL data, along with estimates and credible intervals from the maximal and final Bayesian model.

these two variance components from the model and refitting leads to an identified LMM. Thus, the data of this experiment support six variance components, in agreement with the initial PCA of the maximal model.

Testing non-significant variance components. Given an identified LMM, we test whether removal of any of the remaining variance components reduces the goodness of fit in an LRT. It turns out that all of these variance components are reliable. So we keep them in the model.

Extending the reduced LMM with correlation parameters. Having arrived at an identified reduced LMM, we expand the LMM and check whether there are significant correlation parameters.[3] This model is also supported by the data: There is no evidence of degeneration. Moreover, the model fits significantly better than the zero-correlation parameter model; $\chi^2_{15} = 50, p < .01$. Thus, we would consider this LMM as an acceptable model. The results are documented in Figures 4 and 5.

Pruning low correlation parameters. The significant increase in goodness of fit when going from the reduced zero-correlation parameter model to the extended LMM suggests that there is significant information associated with the ensemble of correlation parameters. Nevertheless, the object and orientation effects and the interaction between spatial and orientation effects are only weakly correlated with the mean as well as with spatial and gravitation effects. So we remove these correlation parameters from the model. There is no loss of goodness of fit associated with dropping most of the correlation parameters; $\chi^2_{12} = 8.4, p = .75$.

Summary The data from this experiment were a follow-up to an experiment reported by Kliegl et al. (2011). The statistical inferences in that article, especially also with respect to correlation parameters, were based on a maximal LMM. A reanalysis along the strategy described here revealed an overparameterization, involving a negative correlation between spatial and attraction effect and a positive correlation between mean and spatial effect. The reanalysis of those early data is also part of the `RePsychLing` package accompanying the present article. The theoretically important negative and positive correlation parameters were replicated with the present experiment in the absence of problems with model complexity (see Kliegl et al. (2015) for further discussion.)

3.4.1 An analysis of Kliegl et al. (2015) using a Bayesian LMM

We also fit a maximal Bayesian linear mixed model using `rstan`. As in the analysis of the Kronmüller and Barr (KB) data, this also showed that the same variance components whose credible intervals do not include zero were the ones that the iterative procedure identified as suitable for inclusion. As in the KB data, notice that the correlation estimates in the Bayesian model tend to be closer to zero than those from `lme4`. This is because the prior on the correlation matrix has most of its probability mass around zero. If the sample size is small, the prior will dominate in determining the posterior distribution of the correlations; but if there is sufficient data, and if a correlation is truly present, the posterior distribution will be different from zero. We see this in the case of three correlation parameters (Figure 5). For comparison, we also show the estimates from a Bayesian model corresponding to the final LMM chosen in the `lme4` analysis presented above.

In sum, the Bayesian analysis independently validates the conclusions based on the PCA-based approach described above.

[3] There is a slight risk that we removed a variance component that would have been significant with correlation parameters in the LMM, but we found no evidence for this in additional analyses.

4 Hidden complexities

Thus far, we have shown that even when models converge, their random effects structure may be overspecified given what the data can actually support. There is another class of problems, however, that an analyst might want to consider. It is well known that any measurement instrument induces changes in what it seeks to measure. Experiments in psychology are no exception. How subjects perform in any given experiment may be co-determined by 'hidden' factors that are not part of the factorial design. Some of these hidden factors can be probed, however. Here, we consider two such factors that are related to the temporal sequence of trials in (behavioral) experiments.

Temporal interdependencies are not part of the simulations which motivate, as an algorithmic proof, the maxim of Barr et al. (2013) to keep random effects structure maximal: Irrespective of model overspecification and failures to converge, their simulations suggest that maximal models best protect against anti-conservative conclusions. Actual empirical data, however, may have further hidden properties that require additional measures on the part of the analyst to ensure an appropriate quantitative assessment.

In what follows, we first return to the `KKL` dataset analysed in the preceding section, to illustrate the presence of these temporal interdependencies, and to illustrate how they can be brought into a mixed effect model. Following this, we briefly discuss an example of a regression analysis in which the same issues are encountered.

4.1 Autocorrelated errors

In sequences of experimental trials, response variables such as reaction times elicited at time t may be correlated with earlier reaction times at $t-k, k \geq 1$, a fact that has been documented across a range of different studies (Broadbent, 1971; Welford, 1980; Sanders, 1998; Taylor and Lupker, 2001; De Vaan et al., 2007; Baayen and Milin, 2010). One source of temporal dependencies between trials is the presence of an autocorrelational process in the errors. Another such source is changes in subjects' predispositions due to effects of learning, fatigue, or modulations in attention. If temporal dependencies are indeed present in the data, they should be brought into the model specification. Failure to do so results in a violation of the assumption that errors are not only identically distributed (sphericity), but also independently distributed. This violation may lead to imprecision in the estimates and their associated test statistics.

The upper panels of Figure 6 illustrate the problem of non-interdependence for the time series of five randomly selected subjects in the `KKL` data. Each panel presents the autocorrelation function for the residuals of the linear mixed model that we fitted to these data. The autocorrelation function graphs the autocorrelation against lag. For lag 0, the autocorrelation is obtained by correlating the vector of trial-ordered reaction times for a given subject with itself. Unsurprisingly, this correlation is 1. At lag 1, this vector is correlated with its values at the preceding point in time $(t-1)$. At lag k, the comparision is with the points at time $t-k$. Figure 6 shows two subjects with minor autocorrelations, a subject with autocorrelations at short lags, and two subjects with autocorrelation that persist across many lags.

For this experiment, a major source of these autocorrelations is slow changes in subjects' attention or concentration over the course of the experiment, potentially confounded with effects of learning or fatigue. In the context of the linear mixed-effects model, one could seek to account for these attentional effects by means of the combination of by-subject random intercepts and by-subject random slopes for trial. However, in our experience, changes in attention tend to reveal a nonlinear functional form. We therefore make use of the generalized additive mixed model (GAMM) (see, e.g., Hastie and Tibshirani, 1990; Lin and Zhang, 1999; Wood, 2006, 2011, 2013), which offers the

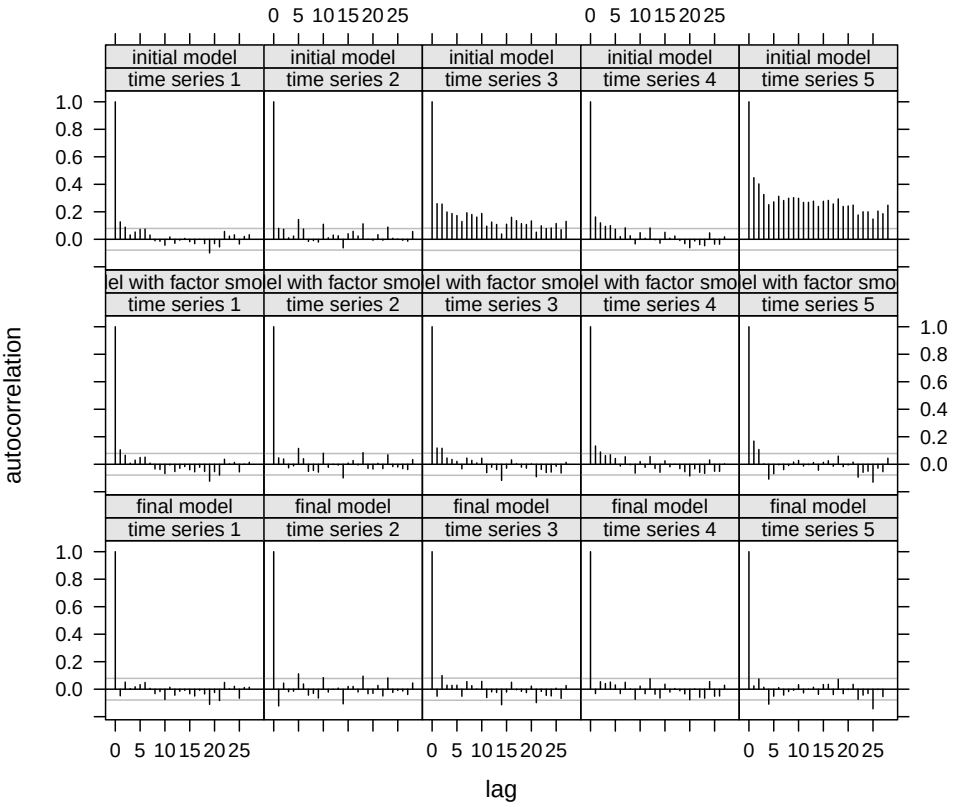

Figure 6: Autocorrelation functions for the residuals of 5 randomly selected subjects' timeseries for the basic mixed model (top panels), a mixed model with by-subject shrunk and penalized factor smooths (second row), and a mixed model that in addition to the factor smooths corrects for AR1 processes in the errors (bottom row). The gray horizontal lines denote approximate 95% confidence intervals.

possibility of accounting for subject-specific wiggly random effects by means of penalized factor smooths which all have the same smoothing parameter. These factor smooths are the non-linear counterpart of the combination of random intercepts and random slopes in the linear model. The second row of panels in Figure 6 clarifies that inclusion in the model of these factor smooths leads to a substantial reduction in the autocorrelation of the errors.

By-subject random wiggly curves estimated with the help of the factor smooths are illustrated in Figure 7. Each curve represents a partial effect for a specific participant, contributed independently of the other effects in the model. The curves show major variation in their vertical positioning, representing by-subject differences in response speed. Most curves show small undulations, which in many cases ride on top of generally downward trends, but for a minority of subjects ride on top of U-shaped trends. It is noteworthy that replacing the factor smooths by random intercepts and random slopes for trial results in a substantial decrease in model fit (to the amount of a loss of 3.4% of variance explained).

Returning to the autocorrelation problem, we note that the panels in the second row of Figure 6 still show some minor autocorrelation, mostly for short lags. These autocorrelations can be accounted for by a simple AR1 autocorrelative process in the errors, according to which the current error is linearly proportional to the preceding error (with a proportionality constant ρ) plus Gaussian noise.

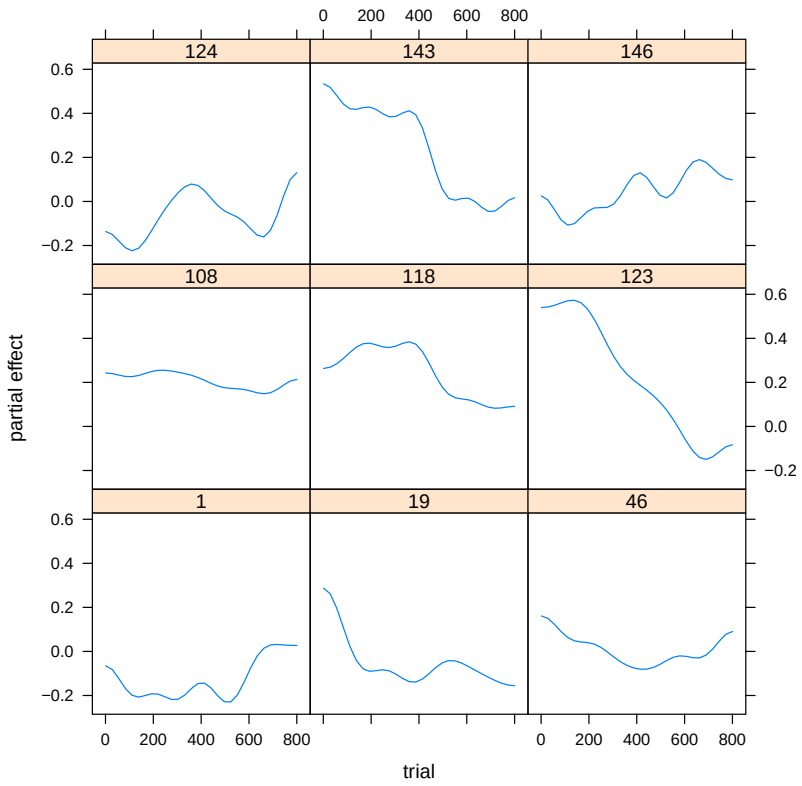

Figure 7: Selected by-subject random wiggly curves in the generalized additive mixed model for the KKL data.

Pinheiro and Bates (2000) and Galecki and Burzykowski (2013) provide extensive discussion of how autocorrelation processes can be accounted for within the mixed modeling framework. At the time of writing, the required algorithms for dealing with autocorrelated errors are not yet available for the `lme4` package. However, the `mgcv` package provides functionality for handling simple AR1 autocorrelations (for tools evaluating AR1 autocorrelation in GAMMS, see also the `itsadug` package, van Rij et al., 2015). With a mild proportionality constant $\rho = 0.15$, the errors become properly uncorrelated, as illustrated by the final row of panels in Figure 6.

4.2 Nonlinearities in control variables

Even when an experiment is designed as factorial, experimental covariates might require further investigation. For the `KKL` data, stimulus onset asynchrony (SOA) was manipulated. SOA is an experimental variable that was included to render unpredictable the moment in time at which a stimulus appeared. Although initially deemed irrelevant, closer scrutiny with a thin plate regression spline revealed a U-shaped effect, as shown in the Figure 8, with optimal response times for SOAs between 400 and 450 ms. Further exploration of this covariate indicated that it did not enter into any interactions with the factorial predictors. It is possible that the effect of SOA varies by subject and trial, but the data are too sparse to allow estimation of such an interaction in the random effects structure.

We note that for data with multiple potentially nonlinear numerical predictors, strategies starting with maximal models are self-defeating. Non-linear interactions can take infinitely many forms.

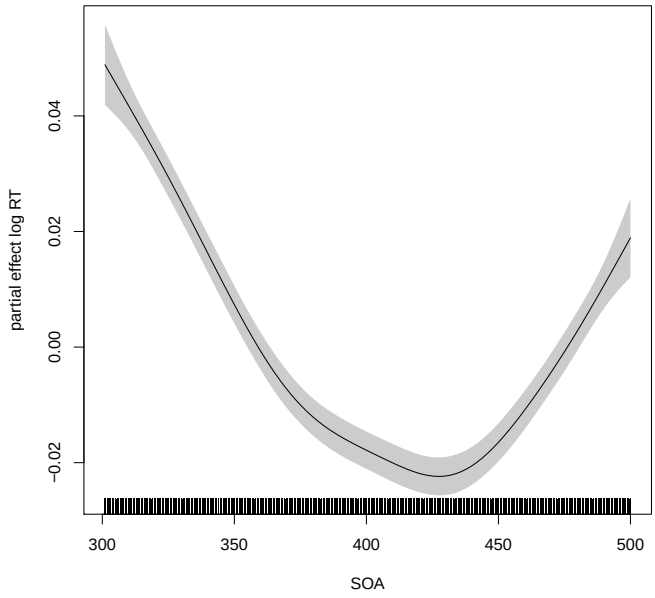

Figure 8: Thin plate regression smooth for SOA.

Without guidance by theory, high-dimension nonlinear surfaces quickly become uninterpretable. Even when tighter fits are obtained, they contribute little to our understanding. As pointed out by Wood (documentation for `gam.selection` in the `mgcv` package),

> The more thought is given to appropriate model structure up front, the more successful model selection is likely to be. Simply starting with a hugely flexible model with 'everything in' and hoping that automatic selection will find the right structure is not often successful.

4.3 Evaluation

It is time to take stock of what we have gained by inspecting the consequences of the experimental procedure itself for our understanding of the forces shaping reaction times in the `KKL` data. First, we have moved from a model that explains 48% of the variance to a model that explains 54% of the variance. Second, and more importantly, the assessment of the factorial predictors for which the experiment was designed changes, especially with respect to the `ord` predictor.

To see this, first consider Table 2, which lists the coefficients (columns 1 and 2) and associated t-statistics (colums 3 and 4) for the initial linear mixed model as well as for the final generalized additive mixed model with by-subject factor smooths, correction for an AR1 process in the errors, and inclusion of a smooth for the SOA covariate.

Many of the fixed-effect coefficients in Table 2 remain very similar, but coefficients where changes are substantial are highlighted. The coefficient for ORN was reduced in magnitude by more than 50%, but remained significant. By contrast, the `sze:orn` interaction increased in magnitude, changing from non-significant to significant. The t-value for the three-way interaction also increased markedly.

As it is difficult to assess how important the changes documented in Table 2 are through inspection of changes in the coefficients, we evaluated variable importance by comparing models with and

	estimate initial	estimate final	t initial	t final
Intercept	5.6910	5.7252	329.4962	322.1204
sze	0.1841	0.1804	5.3305	5.0747
spt	0.0744	0.0728	9.6628	9.2809
obj	0.0409	0.0411	9.0904	10.0563
grv	-0.0015	-0.0005	-0.2862	-0.1059
orn	0.0409	0.0192	3.9167	2.1311
sze:spt	0.0488	0.0480	3.1687	3.0625
sze:obj	-0.0107	-0.0087	-1.1888	-1.0628
sze:grv	-0.0362	-0.0365	-3.3905	-3.7266
sze:orn	0.0165	0.0473	0.7907	2.6270
spt:orn	0.0202	0.0214	3.0219	3.3437
obj:orn	0.0092	0.0081	1.2556	1.1781
grv:orn	0.0110	0.0077	1.4993	1.1289
sze:spt:orn	-0.0127	-0.0099	-0.9478	-0.7727
sze:obj:orn	-0.0019	-0.0080	-0.1327	-0.5836
sze:grv:orn	-0.0435	-0.0486	-2.9539	-3.5441

Table 2: Estimates of the fixed-effects coefficients and associated t-values for the initial model and the final GAMM model (with by-subject random smooths, $\rho = 0.15$, and a thin plate regression spline for SOA as covariate).

without a specific predictor or variance component. If a predictor is important, withholding it will result in a large decrease in goodness of fit. If the predictor is unimportant, withholding it from the model specification should result in little or no change in goodness of fit. We therefore fitted a sequence of models in which terms were withheld from the GAMM model. We assessed the reduction in goodness of fit through the maximum likelihood estimate MLE. The greater the change for the worse in MLE, the greater the variable importance.

More specifically, for the fixed-effect factors, removal of a predictor involved removal of all contrasts and random-effect terms for that predictor. For SOA, only the covariate itself was removed. For `trial`, we replaced the factor smooth by random intercepts for subject, as these are automatically incorporated in the factor smooth for trial. Finally, we assessed the importance of the random-effect terms in the model by removing these terms one at a time. The only random-effect term that was always maintained was that for the by-subject random intercepts (for models with no factor smooth for trial). We followed this procedure first for the initial linear mixed effects model, and then for the final GAMM model. Figure 9 visualizes the resulting variable importance measures by means of a dotplot. Fixed-effect factors and covariates are presented in capitals, and terms for random-effect factors in lower case.

The factor specifying the cue-target relation (with associated dummy variables SPT, OBJ, GRV) has the greatest variable importance (apart from that of the subject random intercepts, not shown). The final model judges this predictor to be even more important than the initial model. The next predictor with the greatest effect on the goodness of fit of the final model is TRIAL. Third in this model comes SOA, followed by the `spt` random-effect contrast. Its variable importance in the final model is similar to that in the initial model. The importance of orientation (ORN) and its interactions in the fixed effect part of the model is much reduced in the final model. The same holds for the random-effect terms involving `orn`. The variable importance of SIZE and the remaining random-effect terms are tiny and comparable across models.

This survey of variable importances indicates that the importance of the cardinal/diagonal

contrast (i.e., the orientation of the stimulus) represented by the ORN and `orn` terms is substantially overvalued by the initial model. The differences between the stimuli involved in this contrast are, apparently, subject to attentional fluctuations and the effects of their processing may linger on to subsequent trials (giving rise to an AR1 process in the errors) to a much greater extent than is the case for the other predictors.

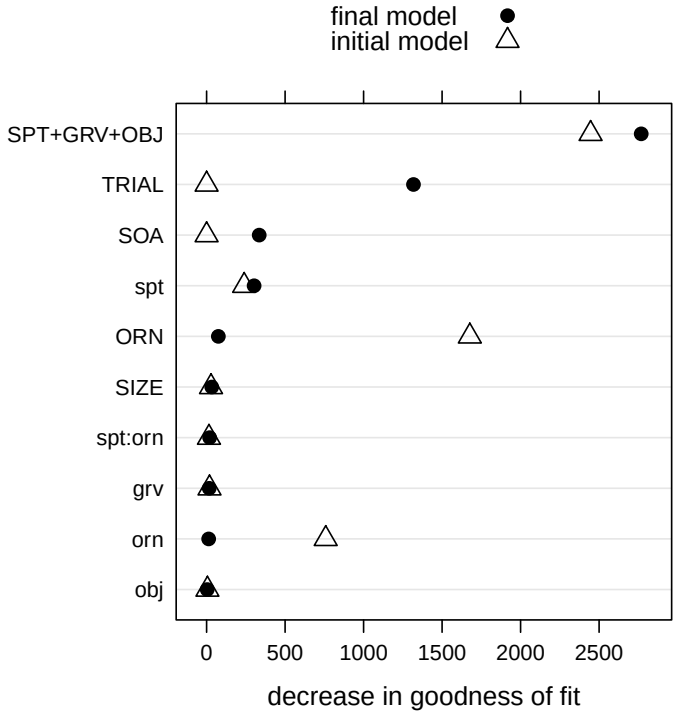

Figure 9: Variable importance for the initial linear mixed model and the final generalized additive mixed model with experimental factors incorporated, evaluated by means of the change for the worse in maximum likelihood when a term is dropped from the model specification. Fixed-effect factors and covariates are in capitals, random effect terms are in lower case.

What this analysis exemplifies is that a factorial experiment analysed with only the factorial predictors in the experimental design may not provide an optimal window into the quantitative structure of the data.

For experiments in physics or chemistry, variability across conditions is much better approximated by random error than is the case for experiments with human subjects, for the simple reason that the experimenter has far better experimental control. When dealing with human subjects, the experimenter is dealing with experimental units that bring their own life with them into the lab. For such experiments, the human factor is worth taking into consideration. Attention may fluctuate, learning may allow for faster responses, and fatigue may slow people down. Some stimuli may be more remarkable than others, resulting in effects that linger on to subsequent trials.

In the presence of non-linear effects due to the human factor, it may become infeasible to implement maximal random effects structure. For the present data, it is conceivable that the effect of SOA changes in the course of the experiment in different ways for different subjects. However, even if it were possible to include multidimensional factor smooths for an interaction of trial by SOA by subject, the present data are not rich enough to support such complex random effects structure.

4.4 Parsimony in regression

Problems related to overparameterization and the human factor are not restricted to factorial designs, but arise also in the context of regression modeling. By way of example, consider the analysis reported by Baayen and Milin (2010) for a reading experiment. The response variable in this study was log-transformed self-paced reading time. The `poems` dataset in the `RePsychLing` package comprises a total of 275996 data points from 326 subjects, for 2315 appearing across 87 modern Dutch poems. Words are partially nested under poems. Any given subject read only a subset of poems. Subject, word, and poem were included as random-effect factors. The original study reported both by-subject and by-word random slopes for several covariates. The correlation parameters for the by-word random slopes, however, were substantial (with absolute magnitudes > 0.8), suggesting overparameterization. Furthermore, Baayen and Milin (2010) sought to eliminate the problem of autocorrelated errors by including trial as a predictor, as well as the latency at the preceding trial. A more principled approach, however, is to include by-subject factor smooths for trial, and probe for an AR1 process for the residual errors.

The problems of overspecification and autocorrelation in regression are well illustrated by considering a small subset of the many predictors considered in the original study: word frequency (`Fre`), participant age in years (`Age`), and participant's reaction times for a multiple-choice question probing their reading habits (`Mul`). Full details are provided in the `gamm` vignette in the `RePsychLing` package. All three predictors were log-transformed and scaled.

A first model requested linear main effects and maximal random effects structure with by-subject random slopes for `Fre`, and by-subject random slopes for `Mul`. Although all variance components were well supported by likelihood ratio tests (all $p < 0.0001$), a principal components analysis of the random-effects variance-covariance estimates indicated that the by-word random slopes (correlation with random intercepts 0.91) account for only 0.14% of the by-word variance-covariance estimates. Inspection of the proportion of variance in the response latencies explained by all terms in the model also indicated that the by-word random slopes have hardly any explanatory value. Since words are partially crossed with subject and with poem, and since the words that appear (repeatedly) in the `poems` dataset are not balanced but follow a Zipfian distribution (as participants were reading continuous text), the data on words are sparse, and not optimal for estimating interactions with subject properties. Therefore, a model without by-word random slopes is a well-motivated alternative to the full model.

For the by-subject random slopes for frequency, the percentage of variance of the subject-specific random-effects variance-covariance estimates captured is an order of magnitude greater (1.32%), but still small in absolute terms. As the parameter for the correlation of the by-subject random slopes and intercepts is smaller in magnitude (-0.68), collinearity is less of an issue. Furthermore, the by-subject random slopes offer a decrease in AIC that is much more substantial than that for the by-word random slopes (2385 versus 74). Data sparsity is also less of an issue.

The negative sign of the correlation parameter is of theoretical interest. Subjects who respond very quickly show little of a frequency effect, whereas subjects who are slow responders reveal a solid effect of frequency. This suggests a potential trade-off between signal-driven responding and responding on the basis of long-term lexical priors.

Further support for the by-subject random slopes for frequency is provided by an interaction of age and frequency (for technical details, see the `RePsychLing` package). Age and frequency are known to enter into interactions in lexical processing (see, e.g., Balota et al., 2004; Ramscar et al., 2014). A nonlinear interaction of `Age` by `Fre` was well supported, and removal of this interaction resulted in a significant decrease in goodness of fit. The interaction indicates that the age effect is slightly smaller for low-frequency words, and that the frequency effect is somewhat stronger for

the younger participants. Interestingly, the by-subject random slopes for `Fre` — which modulate this interaction, albeit indirectly — allow for a more regular and interpretable regression surface to emerge. Removal of the by-subject random slopes appears ill advised, not only quantitatively (as the model fit becomes significantly worse) but also qualitatively, as the regression surface for `Age` by `Fre` becomes more irregular and less well interpretable.

Inspection of the residuals of this model reveals pervasive autocorrelations in the residual errors. Inclusion of a main effect of trial and by-subject random slopes for trial in the linear mixed model only slightly attenuates the autocorrelations. By-subject factor smooths for trial combined with an AR1 process for the errors with proportionality parameter $\rho = 0.3$ eliminated most of the autocorrelation.

In summary, Baayen & Milin (2010) proposed a model with several by-word and by-subject random slopes. Inclusion of these random slopes was motivated in part by the wish to provide stringent tests for the significance of main effects (cf. Barr et al., 2013), and in part by interest in individual differences. Upon closer inspection, the by-word random slopes turned out to contribute very little to the model fit, while at the same time suffering from data sparseness and collinearity. Here, a more parsimonious model without the by-word random slopes seems justified. By contrast, we maintained the by-subject random slopes for frequency. This variance component contributed more substantially to the model fit, and furthermore turns out to be essential for a proper assessment of the interaction of age by frequency. Thus, we kept the model maximal within the boundaries set by what the data can support on the one hand, and by what makes sense theoretically on the other.

5 Discussion

An important goal in statistical analysis of empirical data is the avoidance of overfitting. Any given data set can tolerate only a limited number of parameters. Mixed-effects modeling is no exception. In the statistical literature on fitting mixed-effects modeling (see, e.g. Pinheiro and Bates, 2000; Galecki and Burzykowski, 2013; Bates et al., 2014a), the approach taken is one in which variance components are added to the model step by step, typically driven by theoretical considerations.

The recommendation of Barr et al. (2013) to fit 'maximal' models with all possible random effect components included comes from a very different tradition in which statistics is used to provide a verdict on significance in factorial designs. The authors based their recommendation on a simulation study indicating that anti-conservative results were best avoided by fitting models with as rich a random effects structure as possible.

It is indeed important to make sure that the proper variance components are included in the mixed model. Failure to do so may result in anti-conservative conclusions. However, the advice to "keep it maximal" often creates hopelessly over-specified random effects because the number of correlation parameters to estimate rises quickly with the dimension of the random-effects vectors. The information in the data may not be sufficient to support estimations of such complex models and may result in singular covariance matrices, even when the LMM is identifiable in principle. In this case, we need to replace the complex LMM specification by a more parsimonious one.

With an iterative reduction of the complexity of a degenerate maximal model, one can obtain a model in which the estimated parameters are in line with the information present in the data. We proposed (1) to use PCA to determine the dimensionality of the variance-covariance matrix of the random-effect structure, (2) to initially constrain correlation parameters to zero, especially when an initial attempt to fit a maximal model does not converge, and (3) to drop non-significant variance components and their associated correlation parameters from the model. Each of these reductions may lead to a significant loss in goodness of fit according to LRTs for nested models, in which case this clarifies that the parameter is actually well-supported by information in the data.

Importantly, failure to converge is not due to defects of the estimation algorithm, but is a straightforward consequence of attempting to fit a model that is too complex to be properly supported by the data. We have presented examples showing that the problem of overspecification may arise irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling. Furthermore, even under convergence, overparameterization may lead to uninterpretable models, which is why we developed a diagnostic tool for detecting overparameterization.

What one typically finds for overspecified, degenerate models — degenerate because they afford no predictive power beyond an identified model for the same data — is that the presence of superfluous variance components has minute effects on the estimates of the variability of the fixed-effects estimates. They may occasionally affect standard errors in the decimals, pushing a p-value below or lifting one above some supposedly 'critical' level of α. This should not make a difference as far as conclusions regarding the fixed effects parameters are concerned (Bates et al., 2014a). In fact, comparing parsimonious models with the maximal models discussed by Barr et al. (see the `RePsychLing` package for `R` for full details on the analyses), there is not a single instance where conclusions about the fixed-effect predictors diverge. Thus, for these real data sets, it is not necessary to aim for maximality when the interest is in a confirmatory analysis of factorial contrasts.

If for some reason it is critical to establish the reliability of a specific variance component or correlation parameter, the most promising approach, where feasible, is to collect more data. Beyond that we must mind the Sunset Salvo (Tukey, 1986): "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data" (p.74–75).

What then about the simulation studies on which Barr et al. base their recommendations for maximality? Several issues arise here, which all relate to how representative these simulations are with respect to real data sets. First, the simulations implement a factorial contrast that is atypically large compared to what is found in natural data. Second, and more importantly, the correlations in the random effects structure range from -0.8 to $+0.8$. In our experience, large correlation parameters are diagnostic of overparameterization. They hardly ever represent true correlations in the population. As a consequence, these simulations do not provide a solid foundation for recommendations about how to fit mixed-effects models to empirical data. Third, the simulations presume a pristine world in which the experimental treatments are the only predictors that would be of interest for the confirmatory data analyst. However, experiments run with human subjects have to take into account the human factor. Attention may fluctuate in the course of an experiment, learning takes place, subjects may get tired, and processing may have consequences that linger on and influence subsequent trials. These phenomena are well documented (see, e.g., Broadbent, 1971; Welford, 1980; Sanders, 1998; Taylor and Lupker, 2001; De Vaan et al., 2007; Baayen and Milin, 2010). When during model criticism these kinds of phenomena reveal themselves in the form of autocorrelated errors, inclusion of further predictors such as trial may afford enhanced precision for model parameters and their importance.

In summary, maximal models are not necessary to protect against anti-conservative conclusions. This protection is fully provided by comprehensive models that are guided by realistic expectations about the complexity that the data can support. In statistics, as elsewhere in science, parsimony is a virtue, not a vice.

Acknowledgements

We would like to thank Titus von der Malsburg for detailed comments on an earlier draft.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, U.K.

Baayen, R. H., Davidson, D. J., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Baayen, R. H. and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3:12–28.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., and Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, 133:283–316.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014a). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, page in press.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014b). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Broadbent, D. (1971). *Decision and Stress*. Accademic Press, New York.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

De Vaan, L., Schreuder, R., and Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon*, 2:1–23.

Egly, R., Driver, J., and Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–177.

Forster, K. and Dickinson, R. (1976). More on the language-as-fixed effect: Monte-Carlo estimates of error rates for F_1, F_2, F', and *min*F'. *Journal of Verbal Learning and Verbal Behavior*, 15:135–142.

Galecki, A. and Burzykowski, T. (2013). *Linear mixed-effects models using R. A step-by-step approach*. Springer, New York.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall, London.

Judd, C. M., Westfall, J., and Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, 103(1):54.

Kliegl, R., Kuschela, J., and Laubrock, J. (2015). Object orientation and target size modulate the speed of visual attention. *University of Potsdam*.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., and Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1:1–12.

Kronmüller, E. and Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal of Memory and Language*, 56(3):436–455.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *JRSSB*, 61:381–400.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, R. (2014). Nonlinear dynamics of lifelong learning: the myth of cognitive decline. *Topics in Cognitive Science*, 6:5–42.

Sanders, A. (1998). *Elements of Human Performance: Reaction Processes and Attention in Human Skill*. Lawrence Erlbaum, Mahwah, New Jersey.

Schielzeth, H. and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral ecology*, 20:416–420.

Stan Development Team (2014a). Rstan: the r interface to stan, version 2.5.0.

Stan Development Team (2014b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0*.

Taylor, T. E. and Lupker, S. J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27:117–138.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, 40(1):72–76.

van Rij, J., Baayen, R. H., Wieling, M., and van Rijn, H. (2015). itsadug: Interpreting time series, autocorrelated data using gamms. R package version 0.3 (development version).

Welford, A. (1980). Choice reaction time: Basic concepts. In Welford, A., editor, *Reaction Times*, pages 73–128. Accademic Press, New York.

Westfall, J., Kenny, D. A., and Judd, C. M. (2014). Replicating studies in which samples of participants respond to samples of stimuli. *Journal of experimental psychology: general*, 143(5):2020–2045.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73:3–36.

Wood, S. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100:221–228.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

Zhou, X., Chu, H., Li, X., and Zhan, Y. (2006). Center of mass attracts attention. *Neuroreport*, 17:85—-88.